The road to Superintelligence may have shortened

Tony Czarnecki, Sustensis

London 1/12/2023

Image generated by DALL-E

For an average person, just the term **Artificial Intelligence** (AI) may be quite confusing, as it seems to cover all aspects of what seems to be 'unnatural'. It may start in difficulty to differentiate between Information Technology (IT) and AI.

IT processes information based on strictly defined rules, generally requiring all input data, although there are some heuristic systems that can operate without all data being available. However, AI can produce results based on partially available input data, as it operates similarly to a human mind – using probabilities. It can also learn from experience. Therefore, the same input data may not always produce the same output. The learning experience is what makes some humanoid robots resemble humans – they make errors, but progressively fewer than humans. To make matters even more confusing, many people, including myself, use the term AI as a general descriptor for all types of AI.

What we have now are individual, relatively unsophisticated AI assistants, chatbots such as ChatGPT, or robots. This is generally referred to as Artificial Narrow Intelligence, which is mostly defined as follows:

ARTIFICIAL NARROW INTELLIGENCE (ANI) can exceed human intelligence and capabilities in a single area

These could be games, including poker, which require some intuition, smelling, tasting, or face recognition. ANI can be run on a single computer to perform a single, narrow function supporting one of human skills. However, it is ignorant in all other areas.

By the end of this decade, we may have an **Artificial General Intelligence (AGI)**, which will reach human level intelligence. Wikipedia defines it as "*the ability of an intelligent agent to understand or learn any intellectual task that human beings or other animals can*" [1]. But if we want to build AGI we must have a more detailed definition, identifying its key features. Moreover, we would need to know what that 'intelligent agent' really means, like what I would propose below:

Artificial General Intelligence is a self-learning intelligence, superior to humans', solving any task far better than any human.

How many years away are we from the moment that AI will have human level intelligence making them smarter than humans? Paul Pallaghy, the proponent of Natural Language Understanding theory, who uses a similar definition and predicts AGI may arrive in 2024[1]. I am perhaps a bit more realistic, and like Ray Kurzweil, the renowned futurist, I predict that AGI may emerge by 2030. That prediction is a few decades earlier than many AI researchers still maintain.

If I am right, soon there could be thousands and possibly even millions of AGI humanoids costing perhaps as much as a luxury car. In technology terms it would be a standalone AI system controlling local devices with the access to the Internet. It will need at least these capabilities to achieve a human level intelligence:

- **Short-term memory:** Memorize text, images, graphs and of course events in a conversations (remembering what was said before). OpenAI's GPT-4 Turbo can memorize about 250 pages of text (a whole book), images and graphs similarly as Anthropic's Claude 2.1, so that's done.

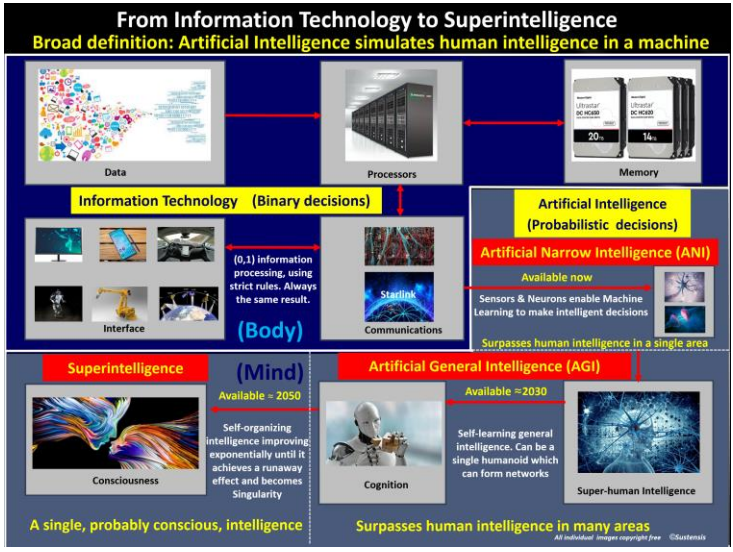
- **Long-term memory:** Record events, topics discussed, and knowledge learned (equivalent to the hippocampus in our brains memorizing events in space and time). That is still limited but should be achieved at an average human level by perhaps the end of the next year, and quite likely be the end of 2025.
- **Multi-step instruction**: Combine intermediate results of individual instructions, building them into the final output. Practically done at a number of companies like Microsoft's Kosmos-1, Google's PaLM-E and several others. It will be perfected by Musk's Optimus and Google's Gemini by the end of 2024 and certainly in 2025, when the first such humanoids will be on sale in limited numbers.
- **Goals and interests:** Create own goals and interests, a kind of a 'free will', which must be compatible with human goals, values, and existing laws - a huge problem of AI Alignment, potentially opening Pandorra's box. We may have to wait till 2027-29.
- **Be truthful and objective**: If AGI is to be human-friendly it must behave following the Universal Values of Humanity. This may require linking goals to human preferences by checking the output. Some progress is being made, e.g., Claude-2 uses its own 'Constitution' to do just that. However, we need to align with an agreed system of Universal Values of Humanity. At the moment top AI developers do it, instead of the World Government. It will be very difficult. If we don't achieve this by 2027-29 and AI gets out of human control, it may potentially become malicious.
- **Emotions:** ChatGPT can detect emotions, and Ameca humanoid can show emotions by following the user's emotions, but they don't feel them. Feeling emotions is not necessary for AGI to have human level intelligence but may be achieved by 2029-30.
- **Cognition:** Simulate human thinking in complex situations, when the answers may be ambiguous or uncertain, using the acquired knowledge, understanding & experience. This is tough but may be achieved comprehensively about 2028-2030.

But the progress can be much faster. Unconfirmed reports indicate that OpenAI was planning to release GPT-5 by the end of 2023, which it describes as a near AGI[2]. The disagreement about the release of such an advanced AI was quite likely the main reason for sacking Sam Altman. If this is the case, OpenAI may have already achieved, what Sam Altman said, a near AGI.

More significantly, in November 2023, NVIDIA released H200 processor, which is many times faster, with much larger memory than H100 supporting all current Large Lange Models (LLM) like GPT-4. Currently, OpenAI uses about 100,000 H100 processors on Microsoft's Azure platform, enabling it to support hundreds of millions users. However, with just one of these processor costing about $40,000, the Meta's Llama-2 can now be run as a standalone version reaching the performance only slightly lower than GPT4. If we extrapolate a near exponential increase in computer power, a standalone computer running AGI could be available by the end of this decade, costing less than a luxury car. The implications of such an early emergence of AGI running on a nearly ubiquitous computer, may fundamentally impact how we live but also how a single person can start a global chaos. That might result when these standalone AGI's become ever more powerful by creating a network and ultimately evolving into **Superintelligence**.

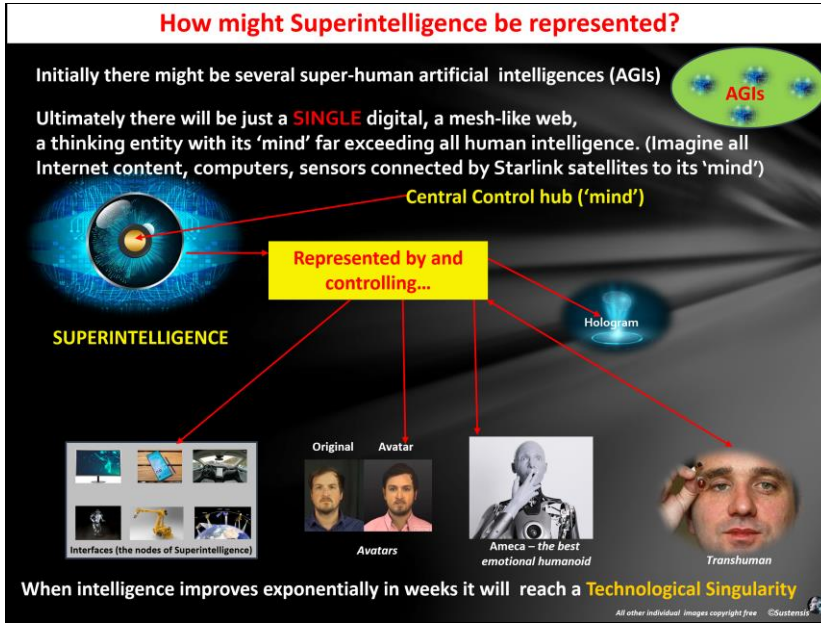
So, what is Superintelligence (or Artificial Superintelligence – **ASI**) and how does it differ from AGI? As with AGI, there is no agreed definition of Superintelligence. I define it as follows:

SUPERINTELLIGENCE is a single self-organizing intelligence, with its own mind and goals exceeding all human intelligence.



Its body consists of various elements such as data, processors, memory, interfaces, communications, sensors, including artificial morphic neurons. All these building blocks are currently thousands of times slower than required for AGI. Therefore, it is unlikely the current AI systems could support AGI with full cognition – an experiential knowledge and awareness of the world. Once it achieves that, it may then gradually turn into a conscious entity. However, there is no agreement among AI researchers whether such an advanced intelligent agent must be conscious before it becomes superintelligent.

Superintelligence will present itself in various ways and through numerous simultaneous representations in any part of the planet [5].



It will appear as avatars, holograms, or as emotional humanoids, such as AMECA robots. It will also be linked to conscious Transhumans, i.e., humans with embedded Brain-Computer-Interface (BCI) via wireless communication with access to external memory and processing power, linking them to AGI and later to Superintelligence.

Superintelligence's behaviour towards humans will depend on whether it has inherited human values, responsibilities, preferences, and expectations. Those values and responsibilities should form a globally agreed Universal Values of Humanity. They should be embedded as early as possible into a top-controlling digital Master Plate of the maturing Superintelligence. However, a mature Superintelligence, millions of times more intelligent than any biological human, will certainly see these values with its own 'Mind' and may thus very quickly replace them with its own. We will not be able to stop it as it develops a separate set of its own values but letting biological humans to govern themselves as much as possible. In this case, even if Superintelligence has a full control over humans, it may not interfere with our lives too much, and instead provide anything we need, creating an unimaginable Global Welfare State. That may be considered an anthropic way of thinking. However, this is similar to humans caring for the animal kingdom (only recently).

If Superintelligence is controlled digitally from 'inside' by Transhuman Governors, then it is highly likely to be human friendly (I cover this subject in my recent book 'Prevail or Fail: A Civilisational Shift to Coexistence with Superintelligence'). It will be seen by biological humans as a single entity, millions of times more intelligent than any human, probably conscious. It may also have billions of complex digital modules, replicating individual human brains, with backup facilities (synchronized copies of the brains). Each such module may be supporting a conscious human mind of a Posthuman. Such modules will likely differ in their capabilities, size, and power to facilitate special roles of certain Posthumans. The Posthumans' 3D representations will be non-biological avatars, probably not conscious but with a high degree of awareness.

If we achieve a full integration with the maturing Superintelligence at a digital level via increasingly more capable BCI devices of the Transhuman Governors, who will ultimately have their brains fully fused (copied) with the 'brain' of Superintelligence, then we may be governed by a Posthuman Government. Should that happen, then it would mean that humans have evolved into digital species. Therefore, there will be no distinction between Superintelligence and Posthumans who will be 'residing' within Superintelligence and being its real mind. In such case, all decisions made by Superintelligence are likely to be

made by a system of voting by all Posthumans and executed by billions of robots and avatars, controlled by Superintelligence.

Some digital Posthumans may be located in space (or their backup copies may be there), for example in Low Earth Orbit (LEO), Geostationary orbit, the Lagrange orbit, on the Moon or even on Mars. However, it is quite likely that in this scenario, vast majority of humans will remain in their biological regenerated bodies for a long time. That means for example, that centenarians may still look and have physical and mental capabilities of biologically much younger people.

Should a full and error-free mind uploading of human brains in a digital form be not possible, then humans will be under a total control of Superintelligence, incapable of understanding the rationale behind some of its decisions. That alone will be an existential threat for humans because we will no longer have any control over our own destiny. Whether such a mature Superintelligence becomes a threat to a human species depends largely on whether it was nurtured in line with human values, so-called AI alignment, before we have completely lost control over it. If Superintelligence has even slightly misaligned objectives or values with those that we share, it may become hostile towards humans.

There are many predictions about the likely time of the emergence of Superintelligence. The date, which is mostly quoted, is 2045, predicted by Ray Kurzweil in his book 'Singularity is Near' published in 2007. He also precited in 2017 that AGI (human-level intelligence) will most likely emerge by 2029. Seeing how surprising were the ChatGPT creators by its vastly better performance than had been expected, and how it has improved over just one year, I would see it as indeed a very likely date. If the speed of AI improvement and its supported hardware continues at the current pace, Ray Kurzweil may also be right about predicting 2045 as the date of the emergence of Superintelligence.

---

[1] Paul Phallaghy: Medium, 25/12/2022, "ChatGPT – The Hard Part of AGI is now done": https://medium.com/@paul.k.pallaghy/chatgpt-the-hard-part-of-agi-is-now-done-3179d31a7277

[2] Eliaçık, Eray "When will GPT 5 be released, and what should you expect from it?" Artificial Intelligence, News, 3/4/2023, https://dataconomy.com/2023/04/chat-gpt5-release-date-agi-meaning-features/